The Science of Health Outcomes Measurement

Neil K. Aaronson, Ph.D.

Head, Division of Psychosocial Research and Epidemiology

The Netherlands Cancer Institute

Until very recently, the conceptual and technical approaches to assessing health outcomes have been grounded in classical psychometric theory and methods. It is only in the past few years that researchers have recognized the potential of modern test theory for advancing the field of health outcomes research, in general, and health-related quality of life (HRQL) research, in particular. The enthusiasm for applying item response theory (IRT) and its pragmatic companion, computer adaptive testing (CAT), in the health and clinical sciences is tangible. There are those who predict that HRQL measurement as we know it today will, in the relatively near future, be relegated to an honorary position as historical artifact. However, others are as yet unconvinced that IRT and CAT will or should supplant more traditional methods, seeing them more as complementary approaches for use in selective situations.

Before we plunge head first into the world of modern test theory, it is perhaps useful to pause for a moment and reflect on what we have (and have not) achieved, to date, in health outcomes and HRQL assessment, and to conjecture about what IRT/CAT may (or may not) have to offer to the field. In this paper, I will present a number of assertions about the current state-of-the-art of HRQL assessment, comment on their tenability and veracity, and speculate on the potential role of IRT/CAT in resolving problem areas and in moving things forward in small steps or in great leaps.

"The term 'health-related quality of life (HRQL)' is well defined and widely understood."

I would argue that this is true if one keeps things simple, but a fiction if one digs a bit deeper. The basic building blocks of HRQL assessment -- physical, mental and social health – are indeed widely accepted and are rooted in literature that goes back to the mid-20th century (e.g., David Karnofsky's definition of "subjective improvement"; the WHO definition of health). However, prevailing definitions of HRQL differ substantially in their focus on health states and functioning, *per se*, versus personal evaluation of those states, in the scope of health issues addressed, and in the polarity of those issues (e.g., dysfunction and its resolution versus well-being). Does it matter? Yes, because the definition of HRQL that we adopt, either implicitly or explicitly, will shape the nature and substance of the questions that we pose regarding the impact of a given health condition and its treatment on the lives of our patients. Can IRT/CAT enhance the conceptual basis of HRQL assessment? Probably not.

"The patient is the sole, legitimate source of information about her HRQL. Other 'proxy' raters (e.g., doctors, nurses, family members) are, at best, poor substitutes."

This, in my opinion, is only a partial truth. Of course, there is nearly universal agreement that the patient is the preferred source of HRQL ratings. However, self-reports can be limited by such factors as age, cognitive status, and symptom distress. In some settings (e.g., pediatrics, psychiatry, geriatrics, palliative care), there may be no alternative to relying on the judgment of others. A recent review of the literature suggests that the use of proxy respondents in selective situations is not only a necessary choice, but also a valid one. Although proxies tend to rate patients as having more problems than do patients themselves, the magnitude of differences in patient-proxy ratings is typically small. However, these findings hold for HRQL measures developed and administered along classical lines. Whether IRT-based assessments, and particularly CAT-versions of HRQL questionnaires will yield the same promising results regarding the appropriateness of proxy assessments is as yet unknown.

Although there are many HRQL questionnaires from which to choose, the dust is settling and a "best bet," can be identified based on a comparison of psychometric characteristics and performance."

This is clearly a fiction, and a slightly dangerous one at that. There is no question that, from the many available generic HRQL questionnaires, the SF-36 has emerged as a market leader. Similarly, within the field of oncology, the EORTC QLQ-C30 and the FACT-G have largely come to dominate the field, albeit with regional and geographic differences in their spheres of influence. However, if one compares the psychometric properties of the larger set of generic and cancer-specific HRQL measures, the differences are relatively small. The empirical literature provides similar levels of support for their underlying measurement models, reliability, validity, responsiveness to change over time, interpretability, respondent burden, and cultural adaptability. IRT-based methods can provide additional insights into the validity (content and construct) of questionnaires, into their cross-language and/or cross-cultural equivalence and, importantly, offer possibilities for calibrating scores across questionnaires so that direct comparison and interpretation of results across studies is possible. Ultimately, however, the choice of instrument to be used in any given study still needs to be driven by other, less technical considerations, among which the specific item content and wording is perhaps one of the most important. My sense is that, with IRT-based methods, we run some risk that the "art" of questionnaire construction, once so central to the development of HRQL measures, will be overwhelmed by purely empirical considerations.

Given the plethora of HRQL questionnaires currently available, there is little or no need for continued efforts at instrument development.

This too is a fiction. The generic HRQL instruments that have been developed on the basis of classical test theory and methods have pretty much reached their performance limits. Any additional gains in precision and efficiency of measurement can probably only be achieved by using the principles and methods of IRT/CAT. Additionally, one of the most important lessons that has emerged from the last decade of HRQL research is that condition-specific questionnaires tend to be more sensitive to group differences and more responsive to intra- and inter-individual changes in health over time. It is for this reason, for example, that both the EORTC and the FACT measurement systems place such a strong emphasis on the development of additional diagnosis-, treatment- and/or symptom-specific questionnaire modules or scales to complement their core measures, and why clinical trial groups such as the Canadian NCI continue to generate symptom checklists for specific treatment trials. Here again, IRT/CAT is likely to play a useful role in the future.

The major methodological challenges in HRQL analysis – missing data, multiple comparisons, and clinical interpretation of statistical results—have been resolved or are well on their way to being resolved.

This statement is reasonably factual. Missing data at the level of individual item responses is a relatively minor and easily resolvable problem. Missing questionnaires, particularly when they reflect systematic loss to follow-up due to illness or death (i.e., informative censoring), represent a more complex problem for which there are workable, albeit imperfect solutions (e.g., mixed effects ANOVA; growth curve analysis). The problem of multiple comparisons and the resulting inflation of p values is inherent to HRQL research. This problem can be minimized, if not eliminated, by using summary scores, when available, by focusing on a few cardinal HRQL outcomes, and/or by applying appropriate statistical adjustments. Defining clinical versus statistical significance is clearly the most difficult and

challenging of these three problems. So-called distribution-based and anchor-based methods are available for translating statistical differences or changes over time into clinically interpretable events. However, no consensus has been reached as to how best to deal with this issue. Most likely, a combination of approaches will be needed. The potential contribution of IRT/CAT to resolving these analytic issues is probably quite limited.

Key stakeholders, including drug regulatory agencies and major clinical trial groups, are increasingly open to and supportive of the use of HRQL outcomes in clinical investigations.

On paper, this statement would appear to be true; in practice, it is rather questionable. For example, as early as 1985, individuals from the U.S.F.D.A. indicated a willingness to consider quality of life outcomes in the oncology drug approval process. Yet, a recent published review indicated that, in the period 1990-2002, no oncology drugs were approved on the basis of HRQL consideration. There were, however, 6 drugs approved (partly) on the basis of symptom relief. In fact, a recent paper by representatives of the Oncology Division of the FDA suggests that the earlier references to quality of life were actually intended to refer primarily and more narrowly to improvement in tumor-related symptoms.

Along similar lines, in the mid-1990's, the American Society of Clinical Oncology (ASCO) identified quality of life as one of the four outcomes of cancer treatment for technology assessment and cancer treatment guidelines. Yet, one of the major contributors to that ASCO document recently coauthored a rather damning review of the added value of HRQL assessments in 46 randomized clinical trials (RCTs) in breast cancer. The authors concluded that HRQL data have contributed meaningfully to RCT's in primary breast cancer, but not to those carried out in the adjuvant setting, in metastatic disease, or in the area of symptom control. They also echoed the FDA position that symptoms, rather than HRQL, should perhaps be the focus of inquiry. The role of HRQL versus symptom-specific data in

clinical trials and in the drug approval and marketing process is the subject of continuing discussion and debate. It is a discussion to which IRT/CAT probably has little to contribute.

Quality of life assessment is ready for prime time as a tool in daily clinical practice.

This statement has an element of truth to it, but may be a bit premature. Contrary to what many believe, the use of HRQL-like measures in clinical practice is not new. As early as 1949, there were reports in the literature on the value of patient-based questionnaire data on physical and psychological symptoms and functioning as an adjunct to a medical interview. More recently, a number of studies have confirmed the feasibility of administering HRQL questionnaires in office-based practice, either in written form or with the use of (touch screen) computers. However, the results of randomized clinical trials that have evaluated the impact of HRQL assessment in daily clinical practice have yielded mixed results. The most consistent positive effect has been on doctor-patient communication, and on physicians' awareness of their patients' problems. Less evidence has accrued regarding the effect on patient management, patient satisfaction with care, and HRQL over time. Importantly, all studies to date have made use of fixed length/fixed format questionnaires designed primarily for use in clinical research (e.g., the EORTC QLQ-C30, the SF-36, the COOP/WONCA charts). Such questionnaires need to be used with caution because of their limited reliability at the level of the individual patient, the absence of clear criteria for defining caseness, and consequently the absence of linkage between HRQL profile scores and treatment/practice guidelines. It is in this area of research and clinical application that IRT/CAT can perhaps contribute most meaningfully. The increased precision and efficiency of IRT-based measures, combined with the dynamic and flexible nature of CAT, holds great promise for facilitating the use of HRQL data as a part of routine clinical practice.

Selective references

Fayers P, Machin D. *Quality of life in clinical trails: Methods and practice*. 2nd ed.. John Wiley & Sons: Chicester, 2000..

Fairclough DL. *Design and analysis of quality of life studies in cliical trials*. Chapman & Hall, CRC Press: Boca Raton,2002.

Goodwin PJ, Black JT, Bordeleau LJ, Ganz PA. Health-related quality of life measurement in randomized clinical trials in breast cancer – taking stock. J Natl Cancer Inst. 2003; 95:263-81.

Greenhalgh J, Long AF, Brettle AJ, Grant MJ. Reviewing and selecting outcome measures for use in routine practice. J Eval Clin Pract. 1998; 4:339-50.

Johnson JR, William G, Pazdur R. Endpoints and Unisted States Food and Drug Administration approval of oncology drugs. J Clin Oncol 2003; 21:1404-1411.

Sneeuw KCA, Sprangers MAG, Aaronson NK. The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: An update. J Clin Epidemiol. 2002; 55:1130-1143.